Hyper-g priors for generalized linear models

Sabanés Bové, D ; Held, L

Abstract: We develop an extension of the classical Zellner's g-prior to generalized linear models. Any continuous proper hyperprior f(g) can be used, giving rise to a large class of hyper-g priors. Connections with the literature are described in detail. A fast and accurate integrated Laplace approximation of the marginal likelihood makes inference in large model spaces feasible. For posterior parameter estimation we propose an effcient and tuning-free Metropolis-Hastings sampler. The methodology is illustrated with variable selection and automatic covariate transformation in the Pima Indians diabetes data set.

DOI: https://doi.org/10.1214/11-BA615

Bayesian Analysis (2011) **6**, Number 3, pp. 387–410

Hyper-g Priors for Generalized Linear Models

Daniel Sabanés Bové* and Leonhard Held†

Abstract. We develop an extension of the classical Zellner's g-prior to generalized linear models. Any continuous proper hyperprior $f(g)$ can be used, giving rise to a large class of hyper-g priors. Connections with the literature are described in detail. A fast and accurate integrated Laplace approximation of the marginal likelihood makes inference in large model spaces feasible. For posterior parameter estimation we propose an efficient and tuning-free Metropolis-Hastings sampler. The methodology is illustrated with variable selection and automatic covariate transformation in the Pima Indians diabetes data set.

Keywords: g-prior, generalized linear model, integrated Laplace approximation, variable selection, fractional polynomials

1 Introduction

Assume that we have observed n independent responses y_i coming from a generalized linear model (GLM, see McCullagh and Nelder 1989) incorporating the covariate vectors $\boldsymbol{x}_i \in \mathbb{R}^p$ via the linear predictors $\eta_i = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}$, $i = 1, \ldots, n$. The response function (inverse link function) h transforms η_i to the mean $\mathbb{E}(y_i) = \mu_i = h(\eta_i)$, which in turn is mapped to the canonical parameter $\theta_i = (db/d\theta)^{-1}(\mu_i)$ of the exponential family. Here $db/d\theta$ is the first derivative of the function b as defined in the likelihood for $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ via

$$f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}) \propto \exp\left\{\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right\}, \tag{1}$$

where each θ_i depends on the intercept β_0 and the vector $\boldsymbol{\beta}$ of regression coefficients as described above. Often the canonical response function $h = db/d\theta$ is used where $\theta_i = \eta_i = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}$. The dispersions $\phi_i = \phi/w_i$ are assumed known and can incorporate weights w_i. The variance $\mathrm{Var}(y_i) = \phi_i d^2b/d\theta^2(\theta_i)$ is expressed through the variance function $v(\mu_i) = d^2b/d\theta^2((db/d\theta)^{-1}(\mu_i))$ as $\mathrm{Var}(y_i) = \phi_i v(\mu_i)$.

A Bayesian analysis starts by assigning prior distributions to the unknown model parameters β_0 and $\boldsymbol{\beta}$. However, usually there is not only uncertainty with respect to the model parameters, but also to the model itself, see e. g. Clyde and George (2004). Let γ be the model index contained in some model space Γ. Typically, the variable selection problem is considered, where $\gamma \in \{0, 1\}^m$ collects binary inclusion indicators for all m available covariates. Here we think more generally of uncertainty about the form (including the dimension p_γ) of the covariate vectors $\boldsymbol{x}_{\gamma i}$, which may also comprise

*Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Switzerland, mailto:daniel.sabanesbove@ifspm.uzh.ch

†Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Switzerland, mailto:leonhard.held@ifspm.uzh.ch

 DOI:10.1214/11-BA615

different transformations of the original variables. For example, when γ indicates a quadratic transformation of x_i, then $\boldsymbol{x}_{\gamma i} = (x_i, x_i^2)^T$. Thus, priors $f(\beta_0, \boldsymbol{\beta}_\gamma \,|\, \gamma)$ need to be assigned, for all models $\gamma \in \Gamma$. Manual elicitation of all these priors is clearly infeasible when Γ is large. In this situation priors which automatically derive from γ are attractive, and we will propose such priors in this paper. Model inference then uses the posterior model probabilities

$$f(\gamma \,|\, \boldsymbol{y}) \propto f(\boldsymbol{y} \,|\, \gamma) f(\gamma), \quad \gamma \in \Gamma, \tag{2}$$

which combine the marginal likelihood

$$f(\boldsymbol{y} \,|\, \gamma) = \int_{\mathbb{R}^{p_\gamma+1}} f(\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}_\gamma, \gamma) f(\beta_0, \boldsymbol{\beta}_\gamma \,|\, \gamma)\, d\beta_0 d\boldsymbol{\beta}_\gamma \tag{3}$$

with the prior model probabilities $f(\gamma)$.

In the special case of the classical normal linear model with known error variance ϕ and $w_i \equiv 1$, the g-prior for the regression coefficients was proposed by Zellner (1986) as a "reference informative prior". It is a mean-zero normal distribution with covariance matrix $g\phi(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}$,

$$\boldsymbol{\beta}_\gamma \,|\, g, \phi \sim \mathrm{N}_{p_\gamma}\big(\mathbf{0}_{p_\gamma}, g\phi(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}\big), \tag{4}$$

and is usually combined with a locally uniform (Jeffreys) prior on β_0, assuming that the design matrix $\boldsymbol{X}_\gamma = (\boldsymbol{x}_{\gamma 1}, \ldots, \boldsymbol{x}_{\gamma n})^T$ has been centered to ensure $\boldsymbol{X}_\gamma^T \mathbf{1}_n = \mathbf{0}_{p_\gamma}$ (see Fernández et al. 2001). Often also the error variance ϕ is assumed unknown and assigned a Jeffreys prior.

The g-prior can be interpreted as the conditional posterior of $\boldsymbol{\beta}_\gamma$ given a locally uniform prior and an imaginary sample $\boldsymbol{y}_0 = \mathbf{0}_n$ from the normal linear model with design matrix $\boldsymbol{X}_\gamma$ and scaled error variance $g\phi$. This reflects the idea that after accounting for the mean level β_0 not included in the g-prior, there is no difference between observations due to the covariates in $\boldsymbol{X}_\gamma$ modelled through $\boldsymbol{\beta}_\gamma$. In addition to this nice interpretation, the g-prior has other advantages, such as invariance of the implied prior for the linear predictor under rescaling and translation of the covariates (Robert and Saleh 1991, p. 71), and automatic adaption to situations with near-collinearity between different covariates (Robert 2001, p. 193).

The hyperparameter $g > 0$ in (4) acts as an inverse relative prior sample size, hence its influence on the results is quite strong. Larger values of g lead to preference of less complex models, a phenomenon known as the Lindley-Jeffreys paradox (Lindley 1957; see also Robert et al. 2009, p. 161). Therefore, much research has been done in developing automatic specifications of g (George and Foster 2000; Hansen and Yu 2001; Fernández et al. 2001; Cui and George 2008). Moreover, a fixed g does not allow the Bayes factor of a perfectly fitting model versus the null model go to infinity (Berger and Pericchi 2001). The multivariate Cauchy priors of Zellner and Siow (1980) correspond to fully Bayesian inference with an inverse-gamma hyperprior for g. Unfortunately, the corresponding marginal likelihood $f(\boldsymbol{y} \,|\, \gamma)$ has no closed form. Therefore Liang

et al. (2008) proposed the hyper-g prior, which is a special case of the incomplete inverse-gamma prior by Cui and George (2008). These hyperpriors retain a closed form expression for $f(\boldsymbol{y} \mid \gamma)$ which is vital for efficient model inference.

In this article we develop an extension of the classical g-prior (4) to GLMs. The hyperprior on the hyperparameter g is handled in a flexible way, so that any continuous proper hyperprior $f(g)$ can be used. In Section 2, this generalized hyper-g prior is derived and connections with the literature are described. Because model inference is the main practical use of this automatic prior formulation, we will propose a fast and accurate numerical approximation of the marginal likelihood in Section 3. Section 3 also covers posterior parameter estimation with a tuning-free Markov chain Monte Carlo (MCMC) sampler. The methodology is applied to variable selection in Section 4 and to fractional polynomial modelling in Section 5. Section 6 discusses possibilities for future research.

2 The generalized hyper-g prior

Section 2.1 derives the generalized hyper-g prior, using arguments analogous to the standard g-prior. Several similar proposals can be found in the literature and are described in Section 2.2.

2.1 Prior construction

Consider the imaginary sample $\boldsymbol{y}_0 = h(0)\mathbf{1}_n$ from the GLM with design matrix $\boldsymbol{X}_\gamma$ (not including an intercept column $\mathbf{1}_n$), original weights vector $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ and scaled dispersion $g\phi$. Using an improper flat prior for the regression coefficients vector $\boldsymbol{\beta}_\gamma$, its posterior given $\boldsymbol{y}_0$ is proportional to the likelihood (1),

$$f(\boldsymbol{\beta}_\gamma \mid \boldsymbol{y}_0, g, \gamma) \propto \exp\left\{ \frac{1}{g\phi} \sum_{i=1}^{n} \Big[h(0) w_i \theta_i - w_i b(\theta_i) \Big] \right\}. \tag{5}$$

This distribution can be recognized as the Chen and Ibrahim (2003, formula 2.6) prior, although the authors have only considered the case $w_i \equiv 1$ and include the intercept β_0. Similar to their theorem 3.1, we can prove that the mode of this distribution is at $\boldsymbol{\beta}_\gamma = \mathbf{0}_{p_\gamma}$ (see the Appendix). It results from standard Bayesian asymptotic theory (e. g. Bernardo and Smith 2000, p. 287) that this distribution converges for $n \to \infty$ to the normal distribution

$$\boldsymbol{\beta}_\gamma \mid g, \gamma \sim \mathrm{N}_{p_\gamma}\big(\mathbf{0}_{p_\gamma}, g\phi c(\boldsymbol{X}_\gamma^T \boldsymbol{W} \boldsymbol{X}_\gamma)^{-1}\big) \tag{6}$$

where $c = v(h(0)) \cdot dh/d\eta(0)^{-2}$ and $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{w})$, because the inverse of the expected Fisher information $I(\boldsymbol{\beta}_\gamma)$ evaluated at the mode is $I(\mathbf{0}_{p_\gamma})^{-1} = g\phi c(\boldsymbol{X}_\gamma^T \boldsymbol{W} \boldsymbol{X}_\gamma)^{-1}$ (cf. Chen and Ibrahim 2003, theorem 2.3).

The "generalized g-prior" (6) differs from the standard g-prior (4) only by the constant c and the weight matrix $\boldsymbol{W}$. Both are especially important in binomial regression

Family	Link	c
Gaussian	Identity	1
	(Log)	1
Poisson	Log	1
	Identity	(0)
Bernoulli	Logit	4
	Cauchit	$\pi^2/4$
	Probit	$\pi/2$
	Complementary log-log	$e-1$
Gamma	Log	1
Inverse-Gaussian	(Log)	1

Table 1: Exponential families, usual link functions and resulting factors c. Note that for the gamma and the inverse-Gaussian family, the natural links μ^{-1} and μ^{-2}, respectively, cannot be used because then $h(0) = \infty$. Parenthesized links should not be used because the uniqueness of the prior mode at $\boldsymbol{\beta}_\gamma = \mathbf{0}_{p_\gamma}$ is not sure (Wedderburn 1976). Parenthesized c's point out problems there.

when w_i is the sample size of the observed proportion, say $y_i = s_i/w_i$ if $s_i \sim \text{Bin}(w_i, \mu_i)$ is the number of successes: In Table 1 it can be seen that only for the Bernoulli family $c \neq 1$. While technically, this scaling constant could be subsumed into the hyperprior on g, it is important because it preserves the interpretation of g as the inverse relative prior sample size, i. e., the prior contains $1/g$ as much information as the data $\boldsymbol{y}$. The use of a common hyperprior $f(g)$ for different exponential families is thus simplified because g always has the same meaning. Although binomial data can always be rephrased as binary data with appropriately replicated covariate vectors and weights $w_i \equiv 1$, this is not possible for non-integer weights w_i where $\boldsymbol{W}$ is absolutely necessary. Non-integer weights are used, for example, for inverse probability weighting (Robins et al. 2000), as sampling weights for survey data (Pfeffermann 1993) and in geographically weighted regression (Brunsdon et al. 1998). Furthermore, note that the g-prior for the normal linear model with independent heteroscedastic errors $\varepsilon_i \sim \text{N}(0, \phi/w_i)$ naturally arises from (6).

Since the intercept β_0 parametrizes the average linear predictor in each model, we can use an improper flat prior $f(\beta_0) \propto 1$. Thus, our generalized g-prior does not shrink the intercept towards zero, while the mean-zero prior on the regression coefficients reflects the idea that $\boldsymbol{X}_\gamma$ has *a priori* no effect on the centered observations. The factor g is assigned a (continuous) hyperprior $f(g)$. In our approach $f(g)$ must be proper to ensure that Bayes factor comparisons with the null model, which does not include the parameter g, are valid. Apart from that, $f(g)$ can be chosen at complete liberty. As g is assigned a hyperprior, we call the resulting prior a "generalized hyper-g prior".

2.2 Comparison with the literature

An immediate question is why we do not use the exact Chen and Ibrahim (2003) prior, which is also a generalization of the standard g-prior. The main problem with this conjugate prior given in (5) is that it does not have a closed form for non-normal exponential families, where the normalizing constant of (5) is unknown. This complicates the computation of the marginal likelihood and the MCMC sampling considerably. Chen et al. (2008) propose a solution where they run an MCMC sampler on the full model, and then derive estimates for submodels. However, this approach is not applicable in problems with simultaneous variable selection and transformation as that presented in Section 5, because no full model exists in that case. Regarding the hyperparameter g, Chen and Ibrahim (2003) propose to assign it an inverse-gamma hyperprior.

Alternatively, Gupta and Ibrahim (2009) proposed the information matrix prior, which uses the expected Fisher information matrix $I(\boldsymbol{\beta}_\gamma)$ similarly to a precision matrix for a normal distribution up to a scalar variance factor g:

$$f_{GI}(\boldsymbol{\beta}_\gamma \mid g, \gamma) \propto |I(\boldsymbol{\beta}_\gamma)|^{1/2} \exp\left\{-\frac{1}{2g}\boldsymbol{\beta}_\gamma^T I(\boldsymbol{\beta}_\gamma)\boldsymbol{\beta}_\gamma\right\}. \tag{7}$$

This will only be a Gaussian distribution if the matrix $I(\boldsymbol{\beta}_\gamma)$ actually does not depend on $\boldsymbol{\beta}_\gamma$, e. g. for the normal linear model where the standard g-prior is reproduced by (7). By contrast, the precision of our generalized g-prior in (6) results from evaluating $I(\boldsymbol{\beta}_\gamma)$ at the prior mode, producing a matrix which does not depend on $\boldsymbol{\beta}_\gamma$. Gupta and Ibrahim (2009) fix the hyperparameter g at a "moderately large" value ($g \geq 1$) and do not consider inference for it.

The information matrix prior is strongly linked with the unit information prior approach of Kass and Wasserman (1995), who proposed the general idea that the amount of information in the prior on $\boldsymbol{\beta}_\gamma$ should be equal to the amount of information about it contained in one observational unit. The amount of information is measured by the (expected) Fisher information, so that the precision is chosen as $n^{-1}I(\mathbf{0}_{p_\gamma})$ in the normal prior

$$f_{KW}(\boldsymbol{\beta}_\gamma \mid g, \gamma) = \mathrm{N}_{p_\gamma}\left(\boldsymbol{\beta}_\gamma \mid \mathbf{0}_{p_\gamma}, nI(\mathbf{0}_{p_\gamma})^{-1}\right). \tag{8}$$

This proposal is close to ours in (6), except that the hyperparameter is fixed at $g = n$. Note that Kass and Wasserman (1995) also required the nuisance parameter β_0 to be (null-)orthogonal to the parameter of interest $\boldsymbol{\beta}_\gamma$, which we ensure by centering the covariates around zero. The unit information prior was used by Ntzoufras et al. (2003) and Overstall and Forster (2010) in the GLM context.

Hansen and Yu (2003, p. 156) also use the expected Fisher information, but evaluate it at the maximum likelihood (ML) estimate $\widehat{\boldsymbol{\beta}}_\gamma$ to obtain a prior precision matrix:

$$f_{HY}(\boldsymbol{\beta}_\gamma \mid g, \gamma) = \mathrm{N}_{p_\gamma}\left(\boldsymbol{\beta}_\gamma \mid \mathbf{0}_{p_\gamma}, gI(\widehat{\boldsymbol{\beta}}_\gamma)^{-1}\right). \tag{9}$$

Hansen and Yu find the dependence of their prior on the data $\boldsymbol{y}$ "hard to accept", although it can be interpreted as an empirical Bayes approach. Also in this flavour, the

authors maximize a cost-modified (approximate) likelihood of g in order to eliminate g. Subsequent model selection is then based on this function value ("minimum description length").

Instead of using the *expected* Fisher information matrix $I(\boldsymbol{\beta}_\gamma)$, Wang and George (2007) use the *observed* Fisher information matrix $J(\boldsymbol{\beta}_\gamma)$. While for canonical response functions the equality $I(\boldsymbol{\beta}_\gamma) = J(\boldsymbol{\beta}_\gamma)$ holds, in general $J(\boldsymbol{\beta}_\gamma)$ is different and depends on the observed response vector. Wang and George (2007) evaluate the observed Fisher information at the original response $\boldsymbol{y}$ and the ML estimate $\widehat{\boldsymbol{\beta}}_\gamma$ to obtain the correlation structure of the normal distribution:

$$f_{WG}(\boldsymbol{\beta}_\gamma \,|\, g, \gamma) = \mathrm{N}_{p_\gamma}\left(\boldsymbol{\beta}_\gamma \,|\, \mathbf{0}_{p_\gamma}, gJ(\widehat{\boldsymbol{\beta}}_\gamma)^{-1}\right). \tag{10}$$

By comparison, our generalized g-prior (6) does not use the original data $\boldsymbol{y}$, but only the design matrix $\boldsymbol{X}_\gamma$. Analogously to Hansen and Yu (2003), Wang and George (2007) select model-specific values for g by maximizing $f(\boldsymbol{y} \,|\, g, \gamma)$, but they also consider fully Bayesian inference for g with flat or truncated-gamma hyperpriors on $1/(g+1)$.

Marin and Robert (2007, p. 101) avoid the use of a Fisher information matrix altogether when they propose the "non-informative g-prior"

$$f_{MR}(\boldsymbol{\beta}_{0\gamma} \,|\, g, \gamma) = \mathrm{N}_{p_\gamma+1}\left(\boldsymbol{\beta}_{0\gamma} \,|\, \mathbf{0}_{p_\gamma+1}, g(\boldsymbol{X}_{0\gamma}^T\boldsymbol{X}_{0\gamma})^{-1}\right) \tag{11}$$

for binary regression with probit and logit link functions, where $\boldsymbol{\beta}_{0\gamma} = (\beta_0, \boldsymbol{\beta}_\gamma^T)^T$ denotes the vector of all coefficients with corresponding full design matrix $\boldsymbol{X}_{0\gamma} = (\mathbf{1}_n, \boldsymbol{X}_\gamma)$. Thus, the intercept β_0 is included in the g-prior. Note that also Gupta and Ibrahim (2009), Hansen and Yu (2003) and Wang and George (2007) originally do not separate the intercept from the other regression coefficients. When $\boldsymbol{X}_\gamma$ is not centered, the intercept is then *a priori* correlated with the other coefficients. In addition, it is also shrunk to its prior mean, not necessarily a desired feature in applications. Marin and Robert (2007) are able to assign g an improper hyperprior, $f(g) \propto g^{-3/4}$, which can be regarded as a degenerate inverse-gamma distribution with shape $-1/4$ and scale 0, because the hyperparameter g is also included in the null model with intercept only.

3 Implementation

In Section 3.1 we propose an accurate numerical approximation of the marginal likelihood under the generalized hyper-g prior. Given a specific model, we can sample from the posterior using a tuning-free Metropolis-Hastings scheme described in Section 3.2. In Section 3.3 we investigate the performance of the numerical and an MCMC marginal likelihood approximation in the conjugate setup, where exact values are known.

3.1 Marginal likelihood computation

Under the generalized hyper-g prior, the marginal likelihood (3) of the GLM γ is

$$\begin{aligned} f(\boldsymbol{y} \mid \gamma) &= \int_{\mathbb{R}^{p_\gamma+1}} f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_\gamma, \gamma) \int_{\mathbb{R}_+} f(\boldsymbol{\beta}_\gamma \mid g, \gamma) f(g)\, dg\, d\beta_0 d\boldsymbol{\beta}_\gamma \\ &= \int_{\mathbb{R}_+} f(\boldsymbol{y} \mid g, \gamma) f(g)\, dg, \end{aligned} \tag{12}$$

where

$$f(\boldsymbol{y} \mid g, \gamma) = \int_{\mathbb{R}^{p_\gamma+1}} f(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}_\gamma, \gamma) f(\boldsymbol{\beta}_\gamma \mid g, \gamma)\, d\beta_0 d\boldsymbol{\beta}_\gamma \tag{13}$$

is the likelihood of g. Note that both (12) and (13) are only defined up to a constant which we have fixed at unity, as we use the improper prior $f(\beta_0) \propto 1$. In general, no closed form expressions are available. The obvious exception is the special case of a Gaussian likelihood, which was mentioned in Section 1 and will be referred to again in Section 3.3. Therefore, in order to be able to efficiently explore a large model space Γ, we need to develop a fast but accurate numerical approximation to the marginal likelihood. This will be a two-step procedure: The likelihood of g in (13) is computed by a Laplace approximation. Plugging this into (12), the hyperparameter g will be integrated out with respect to its prior by numerical integration. Together, this is an integrated Laplace approximation (ILA), which was proposed more generally by Rue et al. (2009).

The Laplace approximation (Lindley 1980; Tierney and Kadane 1986) of (13) is

$$\begin{aligned} f(\boldsymbol{y} \mid g, \gamma) &\approx \frac{f(\boldsymbol{y} \mid \boldsymbol{\beta}^*_{0\gamma}, \gamma) f(\boldsymbol{\beta}^*_{0\gamma} \mid g, \gamma)}{\tilde{f}(\boldsymbol{\beta}^*_{0\gamma} \mid \boldsymbol{y}, g, \gamma)} \\ &= f(\boldsymbol{y} \mid \boldsymbol{\beta}^*_{0\gamma}, \gamma) (2\pi)^{(p+1)/2} \left|\boldsymbol{R}^*_{0\gamma}\right|^{-1/2} \\ &\quad \times (2\pi g\phi c)^{-p/2} \left|\boldsymbol{X}_\gamma^T \boldsymbol{W} \boldsymbol{X}_\gamma\right|^{1/2} \exp\left\{-\frac{1}{2}(g\phi c)^{-1} \boldsymbol{\beta}^{*T}_\gamma \boldsymbol{X}_\gamma^T \boldsymbol{W} \boldsymbol{X}_\gamma \boldsymbol{\beta}^*_\gamma\right\} \end{aligned} \tag{14}$$

when $\tilde{f}(\boldsymbol{\beta}_{0\gamma} \mid \boldsymbol{y}, g, \gamma)$ is the Gaussian approximation of the conditional coefficients posterior with mean vector $\boldsymbol{\beta}^*_{0\gamma}$ and precision matrix $\boldsymbol{R}^*_{0\gamma}$. Since the conditional coefficients prior can be seen to have a normal kernel $f(\boldsymbol{\beta}_{0\gamma} \mid g, \gamma) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T_{0\gamma} \boldsymbol{R}_{0\gamma} \boldsymbol{\beta}_{0\gamma}\right\}$ with (singular) precision

$$\boldsymbol{R}_{0\gamma} = \operatorname{diag}\left\{0, (g\phi c)^{-1} \boldsymbol{X}_\gamma^T \boldsymbol{W} \boldsymbol{X}_\gamma\right\}, \tag{15}$$

the Bayesian iterative weighted least squares (IWLS) algorithm (West 1985; Gamerman 1997) can be used to compute the moments of the Gaussian approximation. Note that this is different and potentially more accurate than the approach by Rue et al. (2009, p. 327) who preserve the sparsity of the prior precision $\boldsymbol{R}_{0\gamma}$ in the resulting posterior precision $\boldsymbol{R}^*_{0\gamma}$. The accuracy of the Laplace approximation (14) can be even further improved by including higher-order terms of the underlying Taylor expansion. For canonical response functions, Raudenbush et al. (2000) derived a convenient correction

factor corresponding to a sixth-order Laplace approximation. In the applications of Sections 4 and 5, we have used this correction (see the Appendix for details), which clearly improved the ILA while requiring only slightly more computation time.

The one-dimensional integration in (12) is performed on the log-scale over $z = \log(g)$ using Gauss-Hermite quadrature. First, we find the (approximate) posterior mode z^* and variance σ^{*2} of z using its unnormalized (approximate) posterior density

$$\tilde{f}(z, \boldsymbol{y} \,|\, \gamma) = \tilde{f}(\boldsymbol{y} \,|\, z, \gamma) f(z). \tag{16}$$

The mode z^* is numerically determined by the `optimize` routine in `R` (R Development Core Team 2010; Brent 1973). The variance σ^{*2} can be computed as the negative inverse second derivative of the log posterior at z^* by numerical differentiation (routine `dfridr` from Press et al. 2007, p. 231). Second, we apply the Gauss-Hermite quadrature (Naylor and Smith 1982)

$$f(\boldsymbol{y} \,|\, \gamma) \approx \sum_{j=1}^{N} m_j \tilde{f}(z_j, \boldsymbol{y} \,|\, \gamma), \tag{17}$$

where the actual weights $m_j = \omega_j \exp(t_j^2)\sqrt{2}\sigma^*$ and nodes $z_j = z^* + \sqrt{2}\sigma^* t_j$ depend on z^*, σ^* as well as original weights ω_j and nodes t_j, $j = 1, \ldots, N$. These can be obtained from the Golub and Welsch (1969) algorithm, which is implemented in the `R`-function `gauss.quad` (Smyth et al. 2010). $N = 20$ seems to be sufficient, given that this includes nodes in a range of about seven standard deviations around z^* (as then $\sqrt{2}t_{20} \approx 7.6$). Note that the Gauss-Hermite approximation in (17) is exact if $\tilde{f}(z, \boldsymbol{y} \,|\, \gamma)$ is the product of $\mathrm{N}(z \,|\, z^*, \sigma^{*2})$ and a polynomial of at most order $2N - 1$.

3.2 Metropolis-Hastings sampler

Given a model $\gamma \in \Gamma$ we would like to sample from the joint posterior of the model-specific parameters $\boldsymbol{\theta}_\gamma = (\boldsymbol{\beta}_{0\gamma}^T, z)^T$. To this end, we propose a tuning-free Metropolis-Hastings (MH) sampling scheme with proposal kernel

$$q(\boldsymbol{\theta}'_\gamma \,|\, \boldsymbol{\theta}_\gamma) = q(\boldsymbol{\beta}'_{0\gamma} \,|\, z', \boldsymbol{\beta}_{0\gamma}) q(z') \tag{18}$$

for the proposal $\boldsymbol{\theta}'_\gamma$ given the current sample $\boldsymbol{\theta}_\gamma$. The independence proposal density $q(z)$ is constructed by first linearly interpolating pairs $\left(z_j, \tilde{f}(z_j, \boldsymbol{y} \,|\, \gamma)\right)$ and second normalizing this function to unity integral, $\int_{\min z_j}^{\max z_j} q(z)\, dz = 1$. Note that many pairs are already available from the optimization and integration of (16) in the marginal likelihood computation, and finer approximations can be obtained by incorporating suitable additional grid points z_j. Thus, $q(z)$ is close to the posterior density $f(z \,|\, \boldsymbol{y}, \gamma)$, suggesting high acceptance rates of the sampler. Also, generating random variates from $q(z)$ using inverse sampling is straightforward as the corresponding cumulative distribution function is piecewise quadratic.

For the coefficients, $q(\boldsymbol{\beta}'_{0\gamma} \,|\, z', \boldsymbol{\beta}_{0\gamma})$ is a Gaussian proposal density: Starting from the current vector $\boldsymbol{\beta}_{0\gamma}$ and the proposed prior covariance factor $g' = \exp(z')$, a single step

of the Bayesian IWLS is made, resulting in the mean vector and the precision matrix of the proposal (Gamerman 1997). In order to compute the acceptance probability of the move from $\boldsymbol{\theta}_\gamma$ to $\boldsymbol{\theta}'_\gamma$,

$$\alpha(\boldsymbol{\theta}'_\gamma \,|\, \boldsymbol{\theta}_\gamma) = 1 \wedge \frac{f(\boldsymbol{y} \,|\, \boldsymbol{\beta}'_{0\gamma}, \gamma) f(\boldsymbol{\theta}'_\gamma \,|\, \gamma)}{f(\boldsymbol{y} \,|\, \boldsymbol{\beta}_{0\gamma}, \gamma) f(\boldsymbol{\theta}_\gamma \,|\, \gamma)} \cdot \frac{q(\boldsymbol{\theta}_\gamma \,|\, \boldsymbol{\theta}'_\gamma)}{q(\boldsymbol{\theta}'_\gamma \,|\, \boldsymbol{\theta}_\gamma)}, \tag{19}$$

note that the prior contributions have the form $f(\boldsymbol{\theta}_\gamma \,|\, \gamma) = f(\boldsymbol{\beta}_\gamma \,|\, g, \gamma) f(g) g$, the last factor g being due to the change of variable $z = \log(g)$ in the proposal parametrization. For the reverse proposal kernel value $q(\boldsymbol{\theta}_\gamma \,|\, \boldsymbol{\theta}'_\gamma)$, another IWLS step starting from the proposed vector $\boldsymbol{\beta}'_{0\gamma}$ and the current factor $g = \exp(z)$ is necessary.

The MH sampler can also be used to compute an MCMC estimate of the marginal likelihood $f(\boldsymbol{y} \,|\, \gamma)$, providing an independent check of the numerical estimate presented in Section 3.1. We will use the method by Chib and Jeliazkov (2001, section 2.1), which was competitive in a review by Han and Carlin (2001) and is still a benchmark for new developments (see e. g. Nott et al. 2008). The estimate is based on the basic identity

$$f(\boldsymbol{y} \,|\, \gamma) = \frac{f(\boldsymbol{y} \,|\, \boldsymbol{\theta}^*_\gamma, \gamma) f(\boldsymbol{\theta}^*_\gamma \,|\, \gamma)}{f(\boldsymbol{\theta}^*_\gamma \,|\, \boldsymbol{y}, \gamma)}, \tag{20}$$

which holds for any $\boldsymbol{\theta}^*_\gamma$. Chib and Jeliazkov (2001) recommend to select $\boldsymbol{\theta}^*_\gamma$ close to the mode of $f(\boldsymbol{\theta}_\gamma \,|\, \boldsymbol{y}, \gamma)$. Detailed balance of the Markov chain ensures that the unknown posterior ordinate can be estimated by

$$f(\boldsymbol{\theta}^*_\gamma \,|\, \boldsymbol{y}, \gamma) \approx \frac{\sum_{j=1}^{B} \alpha(\boldsymbol{\theta}^*_\gamma \,|\, \boldsymbol{\theta}^{(j)}_\gamma) q(\boldsymbol{\theta}^*_\gamma \,|\, \boldsymbol{\theta}^{(j)}_\gamma)}{\sum_{k=1}^{B} \alpha(\boldsymbol{\theta}^{(k)}_\gamma \,|\, \boldsymbol{\theta}^*_\gamma)}, \tag{21}$$

where the $\boldsymbol{\theta}^{(j)}_\gamma$ are the posterior samples and the $\boldsymbol{\theta}^{(k)}_\gamma$ are iid draws from the proposal distribution $q(\boldsymbol{\theta}_\gamma \,|\, \boldsymbol{\theta}^*_\gamma)$. Since each acceptance probability in (21) requires two additional IWLS steps, $4B$ additional IWLS steps are required if B posterior samples are used.

3.3 Performance in the conjugate case

To investigate the performance of the proposed algorithms, we consider the special case of normal linear regression with fixed error variance ϕ. Using the g-prior (4), the conditional coefficients posterior is Gaussian,

$$f(\boldsymbol{\beta}_{0\gamma} \,|\, \boldsymbol{y}, g, \gamma) = \mathrm{N}\left(\beta_0 \,|\, \bar{y}, \phi/n\right) \mathrm{N}_{p_\gamma}\left(\boldsymbol{\beta}_\gamma \,|\, g(g+1)^{-1} \widehat{\boldsymbol{\beta}}_\gamma,\, g(g+1)^{-1} \phi (\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}\right), \tag{22}$$

where the ordinary least squares estimate $\widehat{\boldsymbol{\beta}}_\gamma = (\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1} \boldsymbol{X}_\gamma^T \boldsymbol{y}$ is shrunk by the factor $g/(g+1)$. Thus, the Laplace approximation (14) of the likelihood of g is exact and given by

$$f(\boldsymbol{y} \,|\, g, \gamma) = (g+1)^{-p_\gamma/2} \exp\left\{(g+1)^{-1} \left[-\frac{SSR_\gamma}{2\phi}\right]\right\} \cdot \exp\left\{-\frac{SSE_\gamma}{2\phi}\right\}, \tag{23}$$

where SSE_γ and SSR_γ are the error and regression sums of squares, respectively. From the form of (23) we see that an inverse-gamma hyperprior $\mathrm{IG}(a, b)$ on $g + 1$ will be conjugate to this likelihood. Since $g > 0$ must be ensured, this distribution must be truncated to $(1, \infty)$, yielding the incomplete inverse-gamma prior (Cui and George 2008, p. 891)

$$f(g) = M(a, b)(g + 1)^{-(a+1)} \exp\{-b/(g + 1)\} \tag{24}$$

with normalising constant

$$M(a, b) = \frac{b^a}{\int_0^b t^{a-1} \exp(-t)\, dt} \tag{25}$$

and corresponding marginal likelihood

$$f(\boldsymbol{y} \mid \gamma) = \frac{M(a, b)}{M(a_\gamma, b_\gamma)} \exp\left\{-\frac{SSE_\gamma}{2\phi}\right\}, \tag{26}$$

where the updated parameters $a_\gamma = a + p_\gamma/2$ and $b_\gamma = SSR_\gamma/(2\phi) + b$ determine the posterior of g in model γ.

For illustration, we consider the ozone data introduced by Breiman and Friedman (1985) in the notation of Sabanés Bové and Held (2010), where $n = 330$. Deciding whether to include each of the nine meteorological covariates z_0 and $z_4, \ldots, z_{11}$ in the linear regression of the daily maximum ozone concentration y yields a model space Γ of size $2^9 = 512$. For all $\gamma \in \Gamma$, the ILA (17) and the MCMC estimate (20) of the exact marginal likelihood value (26) were computed fixing the variance at $\phi = 19.75$ (the estimate in the full ordinary linear model) and using the hyperprior parameters $a = 0.01, b = 0.01$. Figure 1 shows that the errors of the ILA and the MCMC estimates are very small here compared to the absolute true values.

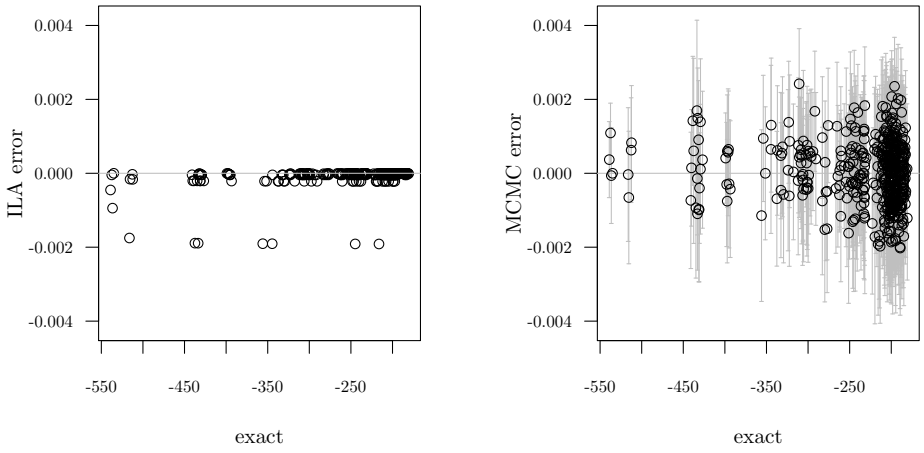
For all models, the acceptance rates of the MH algorithm were above 97%. Figure 2 shows that even for the model with the lowest acceptance rate, the true posterior density of $z = \log(g)$ is very close to its ILA estimate $q(z)$. This explains the almost perfect acceptance rates of the MH scheme.

4 Variable selection

We illustrate the methodology for non-normal data with the Pima Indians diabetes data set (Frank and Asuncion 2010; Ripley 1996), which contains $n = 532$ complete records on diabetes presence and $m = 7$ associated covariates described in Table 2. First, we restrict ourselves to variable selection in the logistic regression model, yielding a model space Γ of size $2^7 = 128$. In Section 5, we will also consider power transformations of the covariates.

Three different hyperprior distributions for the covariance factor g are compared for a fully Bayesian analysis:

F1 $f(g) = \mathrm{IG}(g \mid 1/2, n/2)$, corresponding to the Zellner and Siow (1980) approach;



(a) Errors of the ILA estimates.

(b) Errors of the MCMC estimates. The vertical bars represent 95% MCMC confidence intervals (coverage is 95.1% here).

Figure 1: Errors of the ILA and the MCMC estimates (y-axes) compared to the exact log marginal likelihood values (x-axes) for all 512 models. The MCMC estimates are based on $B = 4500$ samples which were saved after burn-ins of length 1000 (every 2nd iteration). Note that the log marginal likelihood values include the additional additive term $\log \sqrt{2\pi\phi/n}$ compared to (26).

Variable	Description
y	Signs of diabetes according to WHO criteria (Yes = 1, No = 0)
x_1	Number of pregnancies
x_2	Plasma glucose concentration in an oral glucose tolerance test [mg/dl]
x_3	Diastolic blood pressure [mm Hg]
x_4	Triceps skin fold thickness [mm]
x_5	Body mass index (BMI) [kg/m^2]
x_6	Diabetes pedigree function
x_7	Age [years]

Table 2: Description of the variables in the Pima Indians diabetes data set.

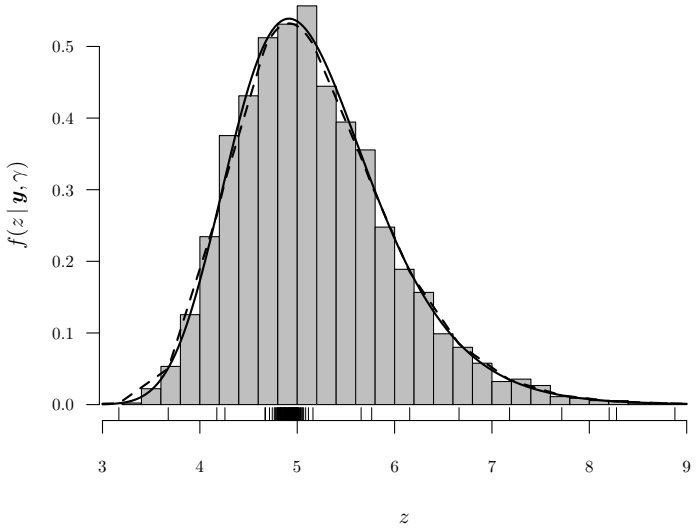


Figure 2: True posterior density of z (solid line) compared with the ILA (dashed line) and MCMC (histogram) estimates. Small ticks above the horizontal axis indicate where nodes z_j for the construction of the ILA estimate $q(z)$ were located (cf. Section 3.2).

F2 $f(g) = 1/n(1 + g/n)^{-2}$, corresponding to the hyper-g/n prior (Liang et al. 2008, p. 416);

F3 $f(g) = \mathrm{IG}(g \mid 0.001, 0.001)$, which is a standard choice for variance parameters.

We also consider model-specific empirical Bayes estimation of g using the likelihood of g in (13), abbreviating this approach as **EB**. Moreover, the standard criteria **AIC** and **BIC** are computed for each model. We use the prior model probabilities

$$f(\gamma) = \frac{1}{m+1}\binom{m}{p_\gamma}^{-1} \tag{27}$$

for an appropriate multiplicity adjustment (George and McCulloch 1993; Scott and Berger 2010). Posterior model probabilities then follow from (2), where for EB the maximized likelihood of g in (13) and for BIC the approximation $\exp(-1/2\,\mathrm{BIC})$ (e. g. Kass and Raftery 1995) is used instead of $f(\boldsymbol{y} \mid \gamma)$. Similar model weights proportional to $\exp(-1/2\,\mathrm{AIC})$ can also be calculated for AIC as proposed by Buckland et al. (1997).

In Table 3, the resulting posterior probabilities and AIC weights for variable inclusion are shown. All methods clearly select x_1, x_2, x_5 and x_6. The corresponding model is the *maximum a posteriori* (MAP) model in F1, F2, F3 and BIC, while for EB and AIC also x_7 is included in the top model. This covariate would be included as well in the median probability model (Barbieri and Berger 2004) for all methods except BIC.

	F1	F2	F3	EB	AIC	BIC
x_1	0.961	0.965	0.968	0.970	0.972	0.946
x_2	1.000	1.000	1.000	1.000	1.000	1.000
x_3	0.252	0.309	0.353	0.384	0.309	0.100
x_4	0.248	0.303	0.346	0.376	0.296	0.103
x_5	0.998	0.998	0.998	0.998	0.998	0.997
x_6	0.994	0.995	0.996	0.996	0.998	0.987
x_7	0.528	0.586	0.629	0.659	0.670	0.334

Table 3: Posterior probabilities and AIC weights for variable inclusion in the Pima Indians diabetes data.

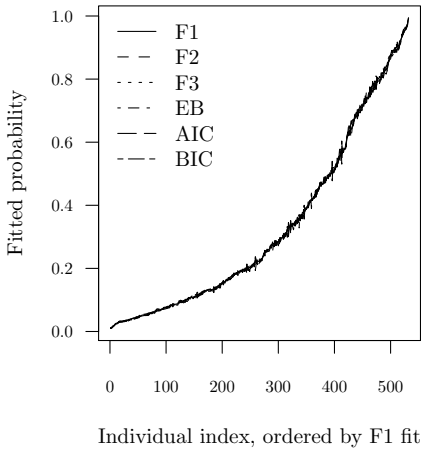
For x_3 and x_4, the evidence for inclusion is consistently weak. For comparison, Holmes and Held (2006) used vague iid normal priors for all coefficients and a flat model prior $f(\gamma) = 2^{-7}$, obtaining clear evidence for inclusion of the MAP covariates.

It is interesting that the inclusion probabilities under F1, F2 and F3 are qualitatively similar. The reason could be that the sample size is relatively large in this example, reducing the importance of the hyperprior specification for g. For EB, most inclusion probabilities are even higher than for F3. The AIC weights are more similar to F2 probabilities (except for x_7). The BIC based probabilities are mostly lower, and close to the (not shown) probabilities under F1 when a flat model prior is used.

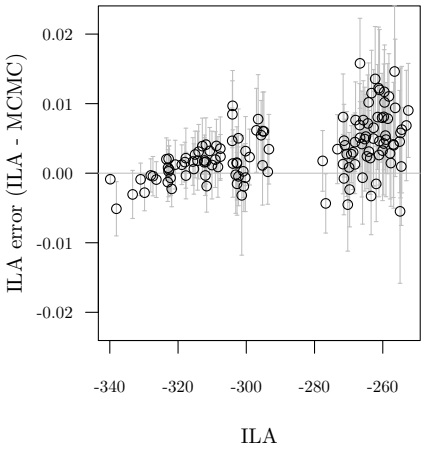
While the posterior inclusion probabilities are visibly different for the six approaches, the model-averaged fits to the data are very close, as shown in Figure 3a. In parallel to sampling the parameters leading to these fitted probabilities for F1, F2, F3 and EB, we also estimated the marginal likelihood by MCMC. The resulting MCMC estimates were close to the ILA estimates, comparison plots looking like Figure 3b for F3. Note that the coverage of the MCMC confidence intervals is lower than in Figure 1b, because the ILA approximations are not exact.

5 Fractional polynomials

Fractional polynomials (FPs) are used for systematic power transformations of the covariates $x_1, \ldots, x_m$ (Royston and Altman 1994). They widen the class of ordinary polynomials insofar as the powers are taken from the fixed set $\{-2, -1, -1/2, 0, 1/2, 1, 2, 3\}$, which also contains square roots, reciprocals and the logarithm by the Box and Tidwell (1962) convention $x^0 \equiv \log(x)$. For each covariate x_k, at most two powers are chosen and collected in the tuple $\boldsymbol{p}_k$, while the corresponding coefficients are collected in the vector $\boldsymbol{\alpha}_k$, determining the FP transform $x_k^{\boldsymbol{p}_k}\boldsymbol{\alpha}_k$. The special case $p_{k1} = p_{k2}$ is handled by multiplication with the logarithm, e. g. $x_k^{(2,2)} = \left(x_k^2, x_k^2 \log(x_k)\right)$. Variable selection is embedded in this framework, because x_k is not included in the model if $\boldsymbol{p}_k = \emptyset$. Each model is thus uniquely identified by $\gamma = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m)$, the covariate vectors are $\boldsymbol{x}_{\gamma i} = (x_{1i}^{\boldsymbol{p}_1}, \ldots, x_{mi}^{\boldsymbol{p}_m})^T$ and the vector of regression coefficients is $\boldsymbol{\beta}_\gamma = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_m^T)^T$.



(a) Model-averaged fitted probabilities.



(b) Errors of the ILA estimates with respect to the MCMC estimates of the log marginal likelihood under F3, for all 128 models. The MCMC estimates are based on (at least) $B = 5000$ samples which were saved after burn-ins of length 1000 (every 2nd iteration). The vertical bars represent 95% MCMC confidence intervals (coverage is 72.7% here).

Figure 3: Results in the Pima Indians variable selection example.

Sabanés Bové and Held (2010) implemented Bayesian model selection for normal linear FP models, and more details on FPs can be found in references therein.

The model space Γ comprises 45^m models, and thus the use of an automatic prior for the parameter $\boldsymbol{\beta}_\gamma$, conditional on the model γ, is very attractive. The generalized g-prior (6) is automatic and only depends on the global hyperparameter g. We will again compare the three fully Bayesian approaches (F1, F2, F3) with the empirical Bayes procedure (EB) which were introduced in Section 4 and avoid manual specification of g. The prior model probabilities $f(\gamma) = \prod_{k=1}^{m} f(\boldsymbol{p}_k)$ depend on the prior FP transformation probabilities

$$f(\boldsymbol{p}_k) = \frac{1}{3}\binom{7 + |\boldsymbol{p}_k|}{|\boldsymbol{p}_k|}^{-1} \tag{28}$$

which have the same form as (27): each degree $|\boldsymbol{p}_k| \in \{0, 1, 2\}$ is equally probable, and all tuples $\boldsymbol{p}_k$ of the same degree are equally probable. This implements Jeffreys's "simplicity postulate" that simpler models must have greater prior probability than more complex models (Jeffreys 1961, section 1.6), and indeed the null model has the largest prior probability 3^{-m}.

For the Pima Indinas diabetes data the model space Γ has size $45^7 \approx 3.7 \cdot 10^{11}$, rendering an exhaustive evaluation of all models $\gamma \in \Gamma$ infeasible. Therefore we use an MCMC model composition (Madigan and York 1995) approach: Starting from the null model, we move through Γ by successive slight modifications of the configuration γ. The modifications are accepted with MH acceptance probabilities, which ensures that models with higher posterior probability are more likely to be visited; see Sabanés Bové and Held (2010) for details. For all four approaches (F1, F2, F3 and EB), we ran this model sampler for one million iterations. To get an idea of the computational complexity, note that on average 10.8 (F2) and 22.1 (EB) models could be evaluated per second (on 2.8 GHz CPUs). All computations have been implemented in an `R`-package including an efficient `C++` core for the MCMC parts, which is available from the first author.

For all four approaches Table 4 shows clear evidence for inclusion of the covariates x_2, x_5, x_6 and x_7 with posterior inclusion probabilities over 99%, while the other three covariates have inclusion probabilities below 15%. In comparison with the variable inclusion results for the untransformed covariates in Table 3, it is interesting that x_1 is no longer important when FP transformations are considered, while x_7 is much more important.

In addition to examining the marginal inclusion probabilities, it is necessary to look at the transformations of the covariates. Since all four approaches produce similar variable inclusion probabilities and also share the MAP model $\boldsymbol{x}_i = (x_{2i}, x_{5i}^{-2}, x_{6i}^{-1/2}, x_{7i}^{-2})^T$, we only look at the F1 approach (the three others give very similar results). In order to account for model uncertainty, it is best to look at model-averaged estimates of variable transformations, conditional on variable inclusion. To this end we varied the transformation of one of the covariates x_2, x_5, x_6, x_7 while fixing the others at their MAP configuration. Averaging over the 44 models each results in the effect estimates shown in Figure 4. Plasma glucose concentration (x_2) seems to have a strong positive linear

	F1	F2	F3	EB
x_1	0.119	0.125	0.135	0.144
x_2	1.000	1.000	1.000	1.000
x_3	0.050	0.052	0.054	0.054
x_4	0.032	0.033	0.033	0.035
x_5	0.999	0.999	0.999	0.999
x_6	0.992	0.993	0.993	0.994
x_7	0.999	0.999	0.999	0.999

Table 4: Posterior probabilities for variable inclusion in the Pima Indians diabetes data when FP transformations are considered. The probabilities are based on 671 525 (F1), 719 929 (F2), 758 616 (F3), and 777 531 (EB) visited models.

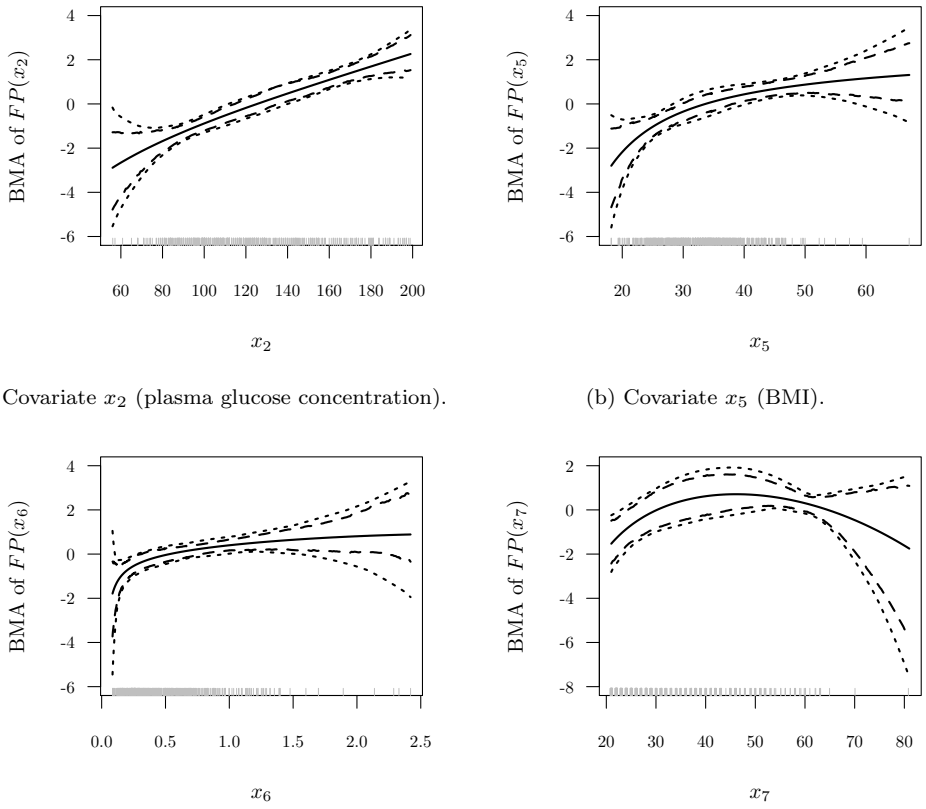
association with diabetes log-odds, while the estimated positive effect of BMI (x_5) is levelling off non-linearly for (rare) high values and is weaker overall. Even smaller is the estimated positive effect of diabetes pedigree function (x_6) with the largest increase in diabetes risk between $x_6 = 0.1$ and $x_6 = 0.5$. The estimated association of age (x_7) is clearly non-linear, with higher diabetes risk for middle-aged participants. These results are qualitatively similar to those obtained by Cottet et al. (2008, p. 665) for a larger subset of the original Pima Indians diabetes data set.

The marginal posterior distributions for the covariance factor g differ slightly between the three hyperprior choices F1, F2 and F3. Averaging over the best 1000 models in terms of posterior probability which have been visited by the model sampler, we get the histograms for $z = \log(g)$ in Figure 5. The corresponding posterior means $\mathbb{E}(g \mid \boldsymbol{y})$ decrease from 282.5 for F1, 219.2 for F2 to 179.1 for F3, and this trend is also visible in the histograms. The results suggest a stronger prior shrinkage of the regression coefficients than that proposed by the unit information prior's fixed value $g = n = 532$ (cf. Section 2.2), as $\mathbb{P}(g < n \mid \boldsymbol{y})$ ranges from 90.9% for F1 to 95.7% for F3.

6 Discussion

In this article, we presented a generalization of the g-prior to GLMs, which can be interpreted analogously to the classical g-prior for normal linear models. In our implementation, the shrinkage-controlling hyperparameter g can be assigned any hyperprior, thus giving rise to a large class of generalized hyper-g priors. For mixtures of classical g-priors, Liang et al. (2008) investigate theoretical model selection and prediction consistency properties. It would be desirable to also investigate such properties for our generalized hyper-g prior class. However, as fewer closed form expressions are available, derivation of comparable proofs will be more difficult in the GLM family.

Another important area of future research is the thorough comparison of the generalized hyper-g prior with the other approaches in the literature summarized in Section 2.2. For example, exhaustive simulation studies could shed light on different performances of



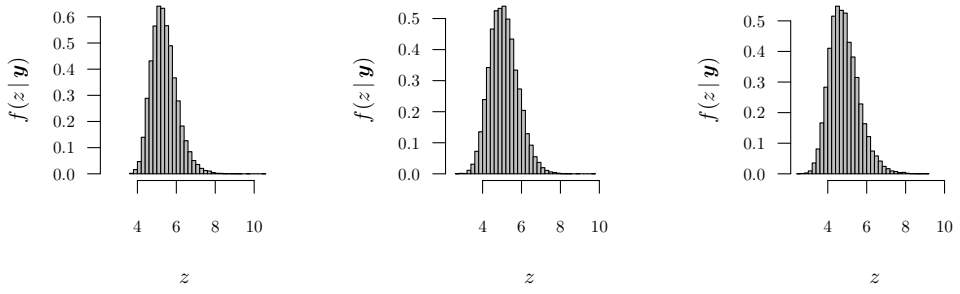
(a) Covariate x_2 (plasma glucose concentration).

(b) Covariate x_5 (BMI).

(c) Covariate x_6 (diabetes pedigree function).

(d) Covariate x_7 (age).

Figure 4: Model-averaged FP transformations of selected Pima Indians covariates under hyperprior F1. Means (solid lines), pointwise (dashed lines) as well as simultaneous (dotted lines) 95% credible intervals are given. Small ticks above the x-axes indicate data locations.



(a) Posterior under hyperprior F1.

(b) Posterior under hyperprior F2.

(c) Posterior under hyperprior F3.

Figure 5: Comparison of marginal posteriors for $z = \log(g)$ under hyperpriors F1, F2 and F3. The histograms are based on the model average over the respective 1000 models with highest posterior probability visited by the model samplers.

the priors in variable selection. Perhaps also theoretical results can be derived to explain the different properties of the approaches. An advantage of our approach is that we allow arbitrary hyperpriors for g while still providing a fast and accurate deterministic approximation to the marginal likelihood.

Bayesian model selection for FPs in GLMs was in fact the motivating application for this work. With huge model spaces to explore, the accurate numerical marginal likelihood approximation is vital for this and similar typical applications of the generalized hyper-g prior. Alternative MCMC estimates of the marginal likelihood were used to demonstrate the very good accuracy of the ILA estimates. Yet, MCMC would not be suited for replacing the deterministic ILA approach in the stochastic model search, because the computation is slower by orders of magnitude and would require careful automatic monitoring of convergence. Of course, the deterministic marginal likelihood approximation could be used for any type of stochastic model search, such as those recently proposed by Hans et al. (2007) and Dobra (2009).

Finally, we note that the classical g-prior has recently been extended in other directions as well. In the context of supervised machine learning, Zhang et al. (2009) replace $\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma$ by a (possibly singular) kernel matrix $\boldsymbol{K}_\gamma$ and prove consistency properties for the normal linear model. Maruyama and George (2010) remove the restriction of $p_\gamma \leq n-1$ for normal linear models by working with the singular value decomposition (SVD) of the design matrix $\boldsymbol{X}_\gamma$. A similar extension is the "generalized singular g-prior" defined by West (2003) in the factor regression context. Along these lines, our generalized hyper-g prior could also be extended to the $p_\gamma > n$ case via the SVD $\boldsymbol{W}^{1/2}\boldsymbol{X}_\gamma = \boldsymbol{U}_\gamma \boldsymbol{D}_\gamma \boldsymbol{V}_\gamma$. We could just use the latent parameter $\boldsymbol{\delta}_\gamma = \boldsymbol{V}_\gamma \boldsymbol{\beta}_\gamma$ of reduced dimension $k_\gamma = n-1$ instead of $\boldsymbol{\beta}_\gamma = \boldsymbol{V}_\gamma^T \boldsymbol{\delta}_\gamma$. Defining the corresponding design matrix as $\boldsymbol{Z}_\gamma = \boldsymbol{W}^{-1/2}\boldsymbol{U}_\gamma \boldsymbol{D}_\gamma$, we have $\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma = \boldsymbol{Z}_\gamma \boldsymbol{\delta}_\gamma$ and retain $\boldsymbol{Z}_\gamma^T \mathbf{1}_n = \mathbf{0}_{k_\gamma}$. Assigning

the prior distribution $\boldsymbol{\delta}_\gamma \sim \mathrm{N}_{k_\gamma}(\mathbf{0}_{k_\gamma}, g\phi c \boldsymbol{D}_\gamma^{-2})$ then induces a normal prior on $\boldsymbol{\beta}_\gamma$ with mean zero and singular precision $(g\phi c)^{-1}\boldsymbol{X}_\gamma^T \boldsymbol{W} \boldsymbol{X}_\gamma$, and thus directly generalizes (6). Investigation of this approach for GLMs with many covariates is another possibility for future research.

Appendix

Proof of prior mode zero

Consider the density function from (5). Dropping for brevity the notational dependency on the model γ, it can be rewritten as

$$f(\boldsymbol{\beta} \,|\, g, \boldsymbol{y}_0) \propto \exp\left\{ \frac{1}{g\phi} \boldsymbol{w}^T \big(h(0)\boldsymbol{\theta} - b(\boldsymbol{\theta}) \big) \right\}, \tag{29}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ and $b(\boldsymbol{\theta}) = \big(b(\theta_1), \ldots, b(\theta_n)\big)^T$. To prove that the mode is at $\boldsymbol{\beta} = \mathbf{0}_p$, note that this is a solution of the score equation

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log f(\boldsymbol{\beta} \,|\, g, \boldsymbol{y}_0) = \frac{1}{g\phi} \left(h(0) \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}^T} - \frac{\partial\, b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}^T} \right)^T \boldsymbol{w} = \mathbf{0}_p,$$

because $\boldsymbol{\beta} = \mathbf{0}_p$ implies that $b'(\theta_i) \equiv b'(\theta) = \mu = h(0)$ and hence

$$\frac{\partial\, b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \mathrm{diag}\big(b'(\theta_1), \ldots, b'(\theta_n)\big) = h(0)\boldsymbol{I}_n.$$

Higher-order Laplace approximation

Denote the standard Laplace approximation (14) by $\tilde{f}_{LA}(\boldsymbol{y} \,|\, g, \gamma)$. Then Raudenbush et al. (2000, p. 148) show that

$$f(\boldsymbol{y} \,|\, g, \gamma) \approx \tilde{f}_{LA}(\boldsymbol{y} \,|\, g, \gamma) \left[1 - \frac{1}{8} \sum_{i=1}^n d_i^{(3)} b_i^2 - \frac{1}{48} \sum_{i=1}^n d_i^{(6)} b_i^3 + \frac{5}{24} \boldsymbol{k}^T (\boldsymbol{R}_{0\gamma}^*)^{-1} \boldsymbol{k} \right] \tag{30}$$

is a sixth-order Laplace approximation when the canonical response function is used. Here $d_i^{(m)} = d^m h / d\eta^m(\eta_i^*)$ evaluated at $\eta_i^* = \boldsymbol{x}_{0\gamma i}^T \boldsymbol{\beta}_{0\gamma}^*$, $b_i = \boldsymbol{x}_{0\gamma i}^T (\boldsymbol{R}_{0\gamma}^*)^{-1} \boldsymbol{x}_{0\gamma i}$ and $\boldsymbol{k} = \sum_{i=1}^n d_i^{(2)} b_i \boldsymbol{x}_{0\gamma i}$. Note that the quadratic forms can be efficiently computed using the Cholesky decomposition $\boldsymbol{R}_{0\gamma}^* = \boldsymbol{L}\boldsymbol{L}^T$, e. g. $\boldsymbol{k}^T (\boldsymbol{R}_{0\gamma}^*)^{-1} \boldsymbol{k} = \|\boldsymbol{v}\|^2$ where $\boldsymbol{L}\boldsymbol{v} = \boldsymbol{k}$.

References

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *Annals of Statistics*, 32(3): 870–897. 398

Berger, J. O. and Pericchi, L. R. (2001). "Objective Bayesian methods for model selection: introduction and comparison." *Lecture Notes-Monograph Series*, 38(1): 135–207. 388

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. Wiley Series in Probability and Statistics. Chichester: John Wiley & Sons. 389

Box, G. E. P. and Tidwell, P. W. (1962). "Transformation of the independent variables." *Technometrics*, 4(4): 531–550. 399

Breiman, L. and Friedman, J. H. (1985). "Estimating optimal transformations for multiple regression and correlation." *Journal of the American Statistical Association*, 80(391): 580–598. 396

Brent, R. P. (1973). *Algorithms for Minimization Without Derivatives*. Prentice-Hall series in automatic computation. Englewood Cliffs, NJ: Prentice-Hall. 394

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998). "Geographically weighted regression—modelling spatial non-stationarity." *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(3): 431–443. 390

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). "Model selection: an integral part of inference." *Biometrics*, 53(2): 603–618. 398

Chen, M. and Ibrahim, J. (2003). "Conjugate priors for generalized linear models." *Statistica Sinica*, 13: 461–476. 389, 391

Chen, M.-H., Huang, L., Ibrahim, J. G., and Kim, S. (2008). "Bayesian variable selection and computation for generalized linear models with conjugate priors." *Bayesian Analysis*, 3(3): 585–614. 391

Chib, S. and Jeliazkov, I. (2001). "Marginal likelihood from the Metropolis-Hastings output." *Journal of the American Statistical Association*, 96(453): 270–281. 395

Clyde, M. and George, E. I. (2004). "Model uncertainty." *Statistical Science*, 19(1): 81–94. 387

Cottet, R., Kohn, R. J., and Nott, D. J. (2008). "Variable selection and model averaging in semiparametric overdispersed generalized linear models." *Journal of the American Statistical Association*, 103(482): 661–671. 402

Cui, W. and George, E. I. (2008). "Empirical Bayes vs. fully Bayes variable selection." *Journal of Statistical Planning and Inference*, 138(4): 888–900. 388, 389, 396

Dobra, A. (2009). "Variable selection and dependency networks for genomewide data." *Biostatistics*, 10(4): 621–639. 404

Fernández, C., Ley, E., and Steel, M. F. J. (2001). "Benchmark priors for Bayesian model averaging." *Journal of Econometrics*, 100(2): 381–427. 388

Frank, A. and Asuncion, A. (2010). *UCI Machine Learning Repository*. URL http://archive.ics.uci.edu/ml 396

Gamerman, D. (1997). "Sampling from the posterior distribution in generalized linear mixed models." *Statistics and Computing*, 7(1): 57–68. 393, 395

George, E. I. and Foster, D. P. (2000). "Calibration and empirical Bayes variable selection." *Biometrika*, 87(4): 731–747. 388

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 398

Golub, G. and Welsch, J. (1969). "Calculation of Gauss quadrature rules." *Mathematics of Computation*, 23(106): 221–230. 394

Gupta, M. and Ibrahim, J. (2009). "An information matrix prior for Bayesian analysis in generalized linear models with high dimensional data." *Statistica Sinica*, 19(4): 1641–1663. 391, 392

Han, C. and Carlin, B. (2001). "Markov chain Monte Carlo methods for computing Bayes factors: A comparative review." *Journal of the American Statistical Association*, 96(455): 1122–1132. 395

Hans, C., Dobra, A., and West, M. (2007). "Shotgun stochastic search for "large p" regression." *Journal of the American Statistical Association*, 102(478): 507–516. 404

Hansen, M. H. and Yu, B. (2001). "Model selection and the principle of minimum description length." *Journal of the American Statistical Association*, 96(454): 746–774. 388

— (2003). "Minimum description length model selection criteria for generalized linear models." *Lecture Notes-Monograph Series*, 40(1): 145–163. Statistics and Science: A Festschrift for Terry Speed. 391, 392

Holmes, C. C. and Held, L. (2006). "Bayesian auxiliary variable models for binary and multinomial regression." *Bayesian Analysis*, 1(1): 145–168. 399

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Oxford University Press, third edition. 401

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90(430): 773–795. 398

Kass, R. E. and Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion." *Journal of the American Statistical Association*, 90(431): 928–934. 391

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481): 410–423. 388, 398, 402

Lindley, D. V. (1957). "A statistical paradox." *Biometrika*, 44(1–2): 187–192. 388

— (1980). "Approximate Bayesian methods." In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, 223–245. Valencia: University of Valencia Press. 393

Madigan, D. and York, J. (1995). "Bayesian graphical models for discrete data." *International Statistical Review*, 63(2): 215–232. 401

Marin, J.-M. and Robert, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer texts in Statistics. New York: Springer. 392

Maruyama, Y. and George, E. I. (2010). "gBF: A Fully Bayes Factor with a Generalized g-prior." Technical report, Center for Spatial Information Science, University of Tokyo.
URL http://arxiv.org/abs/0801.4410 404

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Number 37 in Monographs on Statistics and Applied Probability. Chapman and Hall, second edition. 387

Naylor, J. C. and Smith, A. F. M. (1982). "Applications of a method for the efficient computation of posterior distributions." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3): 214–225. 394

Nott, D. J., Kohn, R. J., and Fielding, M. (2008). "Approximating the marginal likelihood using copula." Technical report, Department of Statistics and Applied Probability, National University of Singapore.
URL http://arxiv.org/abs/0810.5474 395

Ntzoufras, I., Dellaportas, P., and Forster, J. J. (2003). "Bayesian variable and link determination for generalised linear models." *Journal of Statistical Planning and Inference*, 111(1-2): 165–180. 391

Overstall, A. M. and Forster, J. J. (2010). "Default Bayesian model determination methods for generalised linear mixed models." *Computational Statistics and Data Analysis*, 54(12): 3269–3288. 391

Pfeffermann, D. (1993). "The role of sampling weights when modeling survey data." *International Statistical Review*, 61(2): 317–337. 390

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge: Cambridge University Press, 3rd edition. 394

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 394

Raudenbush, S. W., Yang, M.-L., and Yosef, M. (2000). "Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation." *Journal of Computational and Graphical Statistics*, 9(1): 141–157. 393, 405

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press. 396

Robert, C. P. (2001). *The Bayesian Choice*. Springer Texts in Statistics. New York: Springer, second edition. 388

Robert, C. P., Chopin, N., and Rousseau, J. (2009). "Harold Jeffreys's Theory of Probability revisited." *Statistical Science*, 24(2): 141–172. 388

Robert, C. P. and Saleh, A. K. M. E. (1991). "Point estimation and confidence set estimation in a parallelism model: an empirical Bayes approach." *Annales d'Économie et de Statistique*, 23: 65–89. 388

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). "Marginal structural models and causal inference in epidemiology." *Epidemiology*, 11(5): 550–560. 390

Royston, P. and Altman, D. G. (1994). "Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(3): 429–467. 399

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the Royal Statistical Society. Series B (Methodological)*, 71(2): 319–392. 393

Sabanés Bové, D. and Held, L. (2010). "Bayesian fractional polynomials." *Statistics and Computing*. Epub ahead of print, DOI: 10.1007/s11222-010-9170-7. 396, 401

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *Annals of Statistics*, 38(5): 2587–2619. 398

Smyth, G., Hu, Y., and Dunn, P. (2010). *statmod: Statistical Modeling*. R package version 1.4.8. 394

Tierney, L. and Kadane, J. B. (1986). "Accurate approximations for posterior moments and marginal densities." *Journal of the American Statistical Association*, 81(393): 82–86. 393

Wang, X. and George, E. I. (2007). "Adaptive Bayesian criteria in variable selection for generalized linear models." *Statistica Sinica*, 17(2): 667–690. 392

Wedderburn, R. W. M. (1976). "On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models." *Biometrika*, 63(1): 27–32. 390

West, M. (1985). "Generalized linear models: scale parameters, outlier accommodation and prior distributions." In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting*, 531–558. Amsterdam: North-Holland. 393

— (2003). "Bayesian factor regression models in the "large p, small n" paradigm." In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, 733–742. Oxford University Press. 404

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." In Goel, P. K. and Zellner, A. (eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, volume 6 of *Studies in Bayesian Econometrics and Statistics*, chapter 5, 233–243. Amsterdam: North-Holland. 388

Zellner, A. and Siow, A. (1980). "Posterior odds ratios for selected regression hypotheses." In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, 585–603. Valencia: University of Valencia Press. 388, 396

Zhang, Z., Jordan, M. I., and Yeung, D. Y. (2009). "Posterior consistency of the Silverman g-prior in Bayesian model choice." In Koller, D., Bengio, Y., Schuurmans, D., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems (NIPS)*, volume 21. 404

Acknowledgments

The authors would like to thank the referee and the Associate Editor for helpful comments.